DOCUMENT RESUME

ED 466 779 TM 034 290

AUTHOR Hwang, Dae-Yeop

TITLE Classical Test Theory and Item Response Theory: Analytical and Empirical Comparisons.

PUB DATE 2002-02-14

NOTE 27p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Austin, TX, February 14-16, 2002).

PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)

EDRS PRICE EDRS Price MF01/PC02 Plus Postage.

DESCRIPTORS Comparative Analysis; Measurement Techniques; Scores; Statistical Distributions; Test Items; *Test Theory

IDENTIFIERS BILOG Computer Program; Item Characteristic Function

ABSTRACT

This study compared classical test theory (CTT) and item response theory (IRT). The behavior of the item and person statistics derived from these two measurement frameworks was examined analytically and empirically using a data set obtained from BILOG (R. Mislay and D. Block, 1997). The example was a 15-item test with a sample size of 600 examinees (eighth-grade level). The empirical findings indicate that the item and person statistics derived from the two measurement frameworks are quite comparable. The study used a specific characteristic of the test items. Different test score distributions for various item characteristics are recommended for future studies. (Contains 19 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 466 779

Running head: CTT vs. IRT

Classical test theory and item response theory: Analytical and empirical comparisons

Dae-Yeop Hwang

University of North Texas

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
D.-Y. Hwang
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

TM034290

Paper presented at the annual meeting of the Southwest Educational Research Conference, February 14, 2002, Austin, TX.

BEST COPY AVAILABLE

Abstract

This study compared classical test theory (CTT) and item response theory (IRT). The behavior of the item and person statistics derived from these two measurement frameworks was examined analytically and empirically. The empirical findings indicate that the item and person statistics derived from the two measurement frameworks are quite comparable. This study used a specific characteristic of the test items. Different test score distributions for various item characteristics are recommended for future studies.

Classical test theory and item response theory: Analytical and empirical comparisons

Classical test theory (CTT) and item response theory (IRT) have served as two major measurement frameworks for test construction and interpretation. CTT and related models have served test development continuously and successfully over several decades. Recently, the psychometric basis of educational and psychological testing has changed dramatically. IRT has rapidly become mainstream as the theoretical basis for measurement. Increasingly, many standardized tests are developed on the basis of IRT.

Measurement specialists and other test users now have a choice of utilizing CTT and IRT measurement frameworks. The purposes of this paper are (1) to analytically illustrate the depth of the similarities and differences between CTT and IRT and (2) to empirically examine the similarities and differences in the parameters estimated using the two frameworks. This study limits to a simplistic case of IRT models with unidimensionality, dichotomous data, and a one-, two-, and three-parameter models. This study also uses a very simple and easily obtainable dataset for the empirical test.

History of Measurement theories

CTT was pioneered by Spearman (1907, 1913). Gulliksen's (1950) subsequent text is often treated as a classical book for CTT. Traub (1997) highlighted several major concepts in CTT: (1) Correction for attenuation - correlation between variables, (2) Spearman-Brown Prophecy formulas - estimating examinee ability and how the contributions of error might be minimized (e.g., lengthening a test), and (3) Guttman's lower bounds to reliability- reporting true scores or ability scores and associated confidence bands.

Bock (1997) articulated that IRT was initiated by Thurstone (1925). Modern IRT was developed by Lord (1953) and Birnbaum (1957, 1958). Lord and Novick's (1968) classic textbook is considered as a milestone in psychometric methods. Lord and Novick (1968) derived many CTT models from IRT. Rasch (1960), a Danish mathematician, provided a separate line of development in IRT (Embretson & Reise, 2000). Wright further extended Rasch's perspective on latent ability estimation and objective measurement.

The development of psychometric theories and models is related to how to handle measurement errors (Hambleton & Jones, 1993). The specification about error in a model will have substantial impact on how error scores are

estimated and reported (Schumacker, 1998). Under CTT, error might be assumed to be normally distributed. The size of measurement errors might be assumed to be constant across test-score scale (i.e., SEM). However, under IRT, no distributional assumptions about errors are made. The size of errors might be assumed to be related to the examinee's true score. Standard error of measurement is calculated separately for each person measure and each item calibration. If this is the case, more information should result in less error. Embretson and Reise (2000) provided an excellent comparison of CTT and IRT models of measurement analytically and empirically.

Models and Assumptions

Hambleton and Jones (1993) defined the terms "test theories" and "test models". According to their definition, CTT and IRT shall "provide general framework linking observable variables, such as test scores and item scores, to unobservable variables, such as true scores and ability scores." (p. 39). These two test theories are "specified in the form of particular models". Two test models, formulated within the frameworks of the above two test theories, "specify the relationships among a set of test theoretic concepts along with a set of assumptions about the concepts and their relationships."

The CTT model is simple; test scores (often called the observed scores) is the sum of true score and error, $X = T + E$, where X represents the total test score for a particular person, T represents the person's true score on the trait and E represents the person's error on the testing occasion. The above model can be modified into $T = X - E$. Now, true score is defined as the expected test (or observed) score over parallel forms. Parallel forms are defined as tests that measure the same content, have the same true score across persons, and have the equal size of measurement error across forms (Hambleton & Jones, 1993). The resulting two equations are identical and utilized widely in testing practice such as the generalized Spearman-Brown formula, the formula for linking test length to test validity, and disattenuation formulas. Researchers have extended or modified the model within the framework of CTT by dropping or revising one or more of the basic assumptions, or adding distributional assumptions about error and true scores (i.e., the binomial test model).

Test theories and related models provide a framework for practical measurement issues. Different theories and models handle measurement error differently (Hambleton & Jones, 1993, p.39). The assumptions about error for the CTT model are that (a) true scores and error scores are

uncorrelated, (b) the error scores on parallel tests are uncorrelated; the average error score in the population of persons is zero, and (c) error is not correlated with other variables (e.g., true score, other error score and other true scores). Table 1 provides major differences between CTT and IRT.

Insert Table 1 about here

IRT differs substantially from CTT. It is mathematically much more complicated and contains a large family of models. Three frequently used models are one-, two-, and three-parameter IRT models. The following is the most complex three-parameter model (Hambleton & Swaminathan, 1985)

$$P_i(\theta) = c_i + \frac{(1-c_i)e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$$

where c_i is the guessing factor, a_i is the item discrimination parameter (also known as item slope), b_i is the item difficulty parameter (also known as the item location parameter), D is an arbitrary constant, and θ is the ability level of a particular examinee.

This model can be reduced to the one- and two-parameter models if constraints are imposed on two of the three possible item parameters. The three-parameter model is the most general model, and the other two IRT models can be

considered as models nested under the three-parameter model (Hambleton & Swaminathan, 1985).

The one-parameter model is often known as the Rasch model. But, there are fundamental differences between Rasch and the other IRT models (Bode & Wright, 1993). While the Rasch model evaluates the extent to which the data fit its unique definition of measurement based on a stochastic realization of Guttman scaling, IRT searches for any model that will fit whatever data happens to be collected and does not follow the conjoint transitivity recognized by Guttman (Bode & Wright, 1993).

IRT models have two key assumptions: (a) the item characteristic curves (ICCs) have a specified form, and (b) unidimensionality has been obtained (Crocker & Algina, 1993). The general shape of the ICC is specified by a function that relates the person and item parameters to the probabilities (Hambleton & Swaminathan, 1985). Unidimensionality is commonly assumed that only one ability or trait (a single latent ability) is necessary to "explain" or "account" for examinee test performance. The high intercorrelation among test items accounts for by their item parameter (e.g., location, slope etc.) and by their person parameters, as specified in the IRT model.

It does not conflict with the CTT principle of internal consistency (highly correlated items provides more reliable measures) (Hambleton & Swaminathan, 1985).

Test Scores vs. Item Responses

Psychological constructs are conceptualized as latent variables. Latent variables are unobservable entities that influence observable variables such as test scores and item responses (Crocker & Algina, 1986)). Test score or item response is an indicator of a person's standing on the latent variable. Both CTT and IRT provide rationales for behaviorally based measurement. IRT is based on fundamentally different principles than CTT (Embretson & Reise, 2000). IRT is not a mere refinement of CTT; it is a different foundation for testing. IRT provides more complete rationale for model-based measurement than CTT. IRT is a more general foundation for psychological methods.

The CTT model focuses on the test score (or observed score) level. Therefore, the model links test score to true score. True score applies only to a specific set of items on tests with equivalent item properties. Items are regarded as fixed on a particular test. If more than one set of items may measure the same trait, the generality of true score depends on test parallelism or on test equating. These true scores and error scores are not really separable

for an individual score. Instead, the model provides a rationale for estimating true variance and error variance. In CTT, a person's true and error scores cannot be decomposed (Allen & Yen, 1979).

Item properties (i.e., item difficulty and item discrimination) are not explicitly linked to test behavior. Any item properties that are omitted from the model should be justified outside the mathematical model for CTT. The choice of items can be determined by the impact of item difficulty and discrimination on various test statistics, such as variance and reliabilities. In the test development process, both item statistics such as item difficulty (p) and item discrimination (r) and test statistics such as test score mean, standard deviation, and reliability are used to construct tests with the desired statistical properties.

The IRT model links item scores to true scores. The IRT model includes provisions for possibly varying item parameters built in the model. The IRT models include item properties. IRT trait (or ability) levels have meaning for any set of calibrated items. The IRT model can show the relative impact of difficult items on trait level estimates and item responses. In an IRT model, trait (or ability)

level and item properties can be separately estimated (Embretson & Reise, 2000).

CTT involves an additive model. An observed score is the sum of a true score and a random error score. True score and error scores are unobserved constructs. Only observed (or test) scores can be evaluated. Observed scores are computed by summing item scores (0 and 1 for dichotomous or the category numerals in a rating scale). In both dichotomously and polychotomously scored items, the summed scores are treated as linear indicators of the attribute (i.e., higher score indicates more lower score indicates less). But, these observed score sums are neither linear nor equal interval (Wright and Linacre, 1989). In polychotomously scored item (Likert scales), researchers treat the rating scale categories as equal interval and calculate the sum or averages of an item. In CTT, observed scores (called composites) are test dependent; when the items are homogeneous, composites will be high; when the items are not homogeneous, composite will be low.

Under IRT, Rasch weighs the responses by the difficulty levels of the items (Bode & Wright, 1993). Rasch provides estimates of a person's position on a continuum regardless of the difficulty levels of the

particular items asked. IRT focuses on the individual item response rather than the summated test (observed) score as the unit. The Rasch model provides a mathematical procedure for transforming the item responses into measurements with the properties of linearity and specific objectivity (Wright & Masters, 1982). The Rasch model provides a method for examining the item and person order on a single scale continuum, with items and persons serving as the two key factors of the measurement process (Bode & Wright, 1993).

ICC parameters and CTT item statistics

Hambleton and Jones (1993) and Crocker and Algina (1993) showed the Lord (1980)'s mathematical relationship between CTT and IRT. The item-test biserial correlation in CTT and the item discrimination parameter of IRT are approximately increasing functions of each other as follows (Hambleton & Jones, 1993, p. 43)

$$a_i \cong \frac{r_i}{\sqrt{1-r_i^2}}$$

where a_i = item discrimination parameter value for item i for the ICC and r_i = item-total score biserial correlation, which is used as a discrimination index in CTT item analysis. Lord (1980) derived a similar monotonic relationship between the item difficulty parameter of the

ICC, b_i, and the item difficulty estimate for item i, p_i. This monotonic increasing relationship works when all items are equally discriminating (as in the Rasch model). Under this circumstance, as p_i increases, b_i decreases (notice that p_i is an inverse indicator of item difficulty). If all items are not equally discriminating, the relationship between p_i and b_i will depend on r_i. This relationship can be written as (Crocker & Algina, 1986, p. 351)

$$b_i \cong \frac{-\Phi^{-1}(p_i)}{r_i}$$

where p_i is the proportion passing measure of item difficulty for item i, and $\Phi^{-1}(p_i)$ is the z-score of the area p_i to the left of z in the standard normal distribution.

Invariance of item/person statistics.

The most important distinction between CTT and IRT is the property of invariance of both item parameters and ability parameters. Hambleton and Swaminathan (1985) described these two major limitations of CTT and related models.

(a) The item statistic (i.e., item difficulty and item discrimination) is sample (or group) dependent. The p and r values are entirely dependent on the examinee sample from which they are obtained. The higher p values will be obtained from the high ability

sample and the lower p values from the low ability sample. The higher r values will tend to be obtained from heterogeneous examinee sample, and the lower r values from homogeneous examinee samples. The effect of group heterogeneity on correlation coefficients can be found in Lord and Novick (1968).

(b) The person statistic (i.e., test (or observed) score and true scores are test dependent. Consequently, test difficulty directly affects test score or true scores. CTT assumes a very special measurement situation in which examinees are administered the same (or parallel) test items. However, if examinees use several forms of a test with differing difficulty, it is very difficult to compare examinees under the classical test theory. (pp. 1-2)

Two most serious shortcomings of CTT are the sample and test dependences of the person/item statistics. IRT was developed in order to have a test-free and sample-free statistic for dichotomous items. The goal of IRT is to provide both invariant item statistics and ability estimates. In contrast, under the framework of IRT, (a) ability parameters that characterize an examinee are independent of the test items from which they are calibrated and (b) item parameters that characterize an

item are impendent of the ability distribution of a set of examinees (Hambleton & Swaminathan, 1985).

This invariance property of ICCs in the population of examinees for whom the items were calibrated is one of the attractive characteristics of IRT models (Hambleton & Swaminathan, 1985, p.26). The invariance of IRT model parameters has important implications for tailored testing, item banking, item bias, and other applications of IRT (Crocker & Algina, 1986).

Empirical study

The major limitation for CTT is lack of invariance characteristics. CTT does not produce item and person statistics that are invariant across examinee and item samples. The goal of IRT is to provide a test-free and sample-free statistic for dichotomous items. There are just few empirical studies that examine the invariance properties of item statistics from CTT and IRT.

Two studies reported lack of invariance of IRT item parameters (Miller & Linn, 1988; Cook, Eignor, & Taft, 1988). Lawson (1991) examined the comparability of item and person statistics between CTT and Rasch models. He found that person ability estimates and item difficulty estimates were almost identical between two models.

Fan (1998) replicated the study by Lawson (1991) with a large-scale state assessment database. His empirical study focused on two major issues: (a) The comparability of the item and person statistics between CTT and IRT and (b) the invariance characteristics of the item statistics between CTT and IRT across examinee samples. Similar to Lawson (1991), he found that the person and item statistics derived from the two frameworks were quite comparable, and the degree of item statistics across samples also appeared to be similar for the two measurement frameworks.

In the present empirical study, a data set was obtained from BILOG (Mislevy & Bock, 1997) Example 6 consisting of a fifteen-item test from a test of mathematics at the eight-grade level. A sample of size 600 was randomly selected from the data file for the purpose of the calibration. This empirical study only focuses on the comparability of CTT and IRT item statistics. The comparability of CTT- and IRT- based item statistics was examined by correlating CTT and IRT item statistics obtained from a sample. Two types of item statistics were compared: (a) item difficulty parameter b from IRT models with CTT item difficulty p value and (b) IRT item discrimination parameter a (item slope parameter from two- and three-parameter IRT models) with CTT item

discrimination index (item-test, point-biserial correlation).

ITEMAN Version 3.6 (1998), RASCAL Version 3.0 (1997), BILOG Version 3.11 (1997) were utilized for this empirical study under the frameworks of CTT, Rasch and IRT. For CTT, item statistics (i.e., total test ability scores, item difficulty and item-total point-biserial correlation coefficients) were computed. Rasch statistics (i.e., person ability estimates and item difficulty estimates) were obtained from Rascal. Item statistics from one-, two-, and three-parameter models were obtained through the use of BILOG Version 3.11 (1997). The three-parameter IRT model was used for the multiple-choice items.

Results of the CTT, Rasch, and IRT models for the data set are presented in Table 2 through 5. The first two columns of Table 2 represent estimates of individual abilities as reflected by the number of correct item responses. Column 2 in Table 2 presents person ability estimates provided through the Rasch procedure. Column 3 in Table 2 indicates the item numbers from the item pool that were used to calculate the estimates of both item difficulty and item discrimination. All the three models'

Insert Tables 2-5 about here

---

difficulty estimates are presented in the next five columns. The last three columns in Table 2 represent estimates of each item's ability to discriminate between ability levels of examinees. Tables 3 through 5 provide Pearson product-moment correlations obtained from each model to investigate comparability of CTT and IRT item statistics.

Conclusion

The present study compared two measurement theories analytically and empirically. Analytically, IRT is a more robust measurement method. It can produce a test-free and sample-free statistics for dichotomous items. However, empirically, the results did not justify the difference between the two methods.

As in Lawson (1991) and Fan (1998), the correlation coefficients found in this study indicate that there are considerable similarities between the item statistics obtained through CTT and IRT. Both procedures produce almost identical information regarding both item difficulties and item discriminations.

However, this finding does not necessarily discredit the applicability of IRT model procedures. Lawson (1991) and Fan (1998) recognized the limitations of their empirical studies. Fan suggested two major limitations

regarding the data: (1) the characteristics of the test items and (2) limited item pool used in his empirical study. In particular, the test score distribution in the Fan's (1998) study had strong ceiling effect. The strong ceiling effects suggest that many items tended to be very easy. As in his study, the present study uses a very specific characteristic of the test items. In future study, the test item pool should be larger and more diverse so that items can be sampled from the pool under different conditions of item characteristics (Fan, 1998, p. 379). Future studies should use items varying more in item difficulty and in item discrimination. We can use various test score distributions such as negatively skewed, positively skewed, or bimodal distributions.

Two decades ago, Robert L. Thorndike (1982) summed up the future of IRT models

> For the large bulk of testing, both with locally developed and with standardized tests, I doubt that there will be a great deal of change. The items that we will select for a test will not be much different from those we would have selected with earlier procedures, and the resulting tests will continue to have much the same properties.

If this is the case, one must ask, "so much work for so little gain?"

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. CA: Brooks/Cole Publishing Company.

Bock, D. R. (1997). A brief history of item response theory. *Educational Measurement: Issues and Practice*, 21-33.

Bode, R. K., & Wright, B. D. (1993). Rasch measurement in higher education. In B. (Ed.), *Research in Higher Education*, Volume 9, 287-316, New York, NY: Basic Books.

Cook, L L., Eignor, D. R., & Taft, H. L. (1988). A comparative study of the effects of recency of instruction on the stability of IRT and conventional item parameter estimates. *Journal of Educational Measurement*, 25, 31-45.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehart & Winston.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. New Jersey: Lawrence Erlbaum Associates.

Fan, X. (1998). Item response theory and classical test

theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement*, 58 (3), 357-381.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational Measurement*, Third Edition. New York: Macmillan Publishing.

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their application to test development. *Educational Measurement: Issures and Practice*, 12(3), 38-47.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and Applications*. Boston: Kluwer.

ITEMAN (Version 3.6)(1998). St. Paul, MN: assessment systems Corporation.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments*, 1, 159-168, Greenwich, CT: JAI.

Miller, M. D., & Linn, R. L. (1988). Invariance of item characteristic functions with variations in instructional coverage. *Journal of Educational Measurement*, 25, 205-219.

Mislevy, R. J., & Bock, D. R. (1997). BILOG (Version 3.11). Lincolnwood, IL: Scientific software International.

RASCAL (Version 3.0)(1997). St. Paul, MN: Assessment systems Corporation.

Schumacker, R. E. (1998). Comparing measurement theories. Paper presented at the Annual Meeting of the American Educational Research Association (San Diego, CA, April 13-17, 1998).

Traub, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 8-14.

Wright, B. D., & Lincre, J. M. (1989). Observations are always ordinal; measurements, however, must be interval. *Archives of Physical Medicine and Rehabilitation*, 70 (12), 857-860.

Wright, B. D., & Master, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Table 1
Main differences between CTT and IRT

	CTT	IRT
Model	Linear $X = T + E$	Nonlinear $P_i(\theta) = c_i + \frac{(1+c_i)e^{Da_i(\theta-b_i)}}{1+e^{Da_i(\theta-b_i)}}$
Assumptions	Weak (i.e., easy to meet with test data) • $E(e) = 0$ • $\rho_{TE} = 0$ • $\rho_{e_1e_2} = 0$	Strong (i.e., more difficult to meet with test data) • Unidimensionality (dependence among items or number of latent traits needed to achieve local independence) • Local independence (independece among items at ability levels)
Level	Test	Item
Error of Measurement	Error = X - T	Error=Observed-Predicted Response Response
Score Interpretation	X + SEM	Rasch: logit ± residual IRT: θ ± error where score indicates probability of responding correctly to an item given latent model
Item-ability Relationship	Not specified	ICC
Item statistics	p, r	a, b, c (for the 3-parameter model)
Ability	Test scores (or estimated true scores) are reported on the test-score scale)	Ability scores are reported on the scale $-\infty$ to $+\infty$
Invariance of Item & Person	No - item & person parameters are sample dependent.	Yes - item & person parameters are sample independent, if model

statistics		fits the test data. • Test-free measurement • Sample-free measurement
Sample Size	200 to 500 (in general)	Depends on the IRT model but larger samples (over 500), in general, are needed.

Table 2
Comparability of Ability and Item Statistics from the Two Measurement Frameworks

Person Ability		Item Difficulty					Discrimination		
N	Rasch	No.	CTT	Rasch	1P	2P	3P	CTT	2P	3P
1[a]	-3.95	1	.838[b]	-2.052	-1.634	-1.822	-1.631	.37[c]	.659	.699
2	-2.68	2	.970	-4.173	-3.199	-4.886	-4.169	.16	.453	.530
3	-1.90	3	.678	-0.969	-0.777	-0.862	-0.690	.47	.661	.701
4	-1.33	4	.488	0.120	0.020	0.017	0.490	.56	.706	.302
5	-0.87	5	.587	-0.377	-0.382	-0.372	-0.236	.58	.863	.910
6	-0.46	6	.535	-0.076	-0.169	-0.166	0.004	.60	.980	1.140
7	-0.08	7	.497	0.141	-0.013	-0.045	0.055	.68	1.570	1.708
8	0.29	8	.560	-0.245	-0.272	-0.324	-0.049	.50	.585	.680
9	0.67	9	.627	-0.545	-0.550	-0.474	-0.344	.63	1.139	1.222
10	1.07	10	.390	0.700	0.428	0.396	0.516	.57	.851	1.000
11	1.50	11	.453	0.360	0.164	0.097	0.175	.69	1.460	1.584
12	2.01	12	.358	0.746	0.566	0.605	0.869	.56	.687	1.458
13	2.63	13	.183	2.162	1.467	1.321	1.327	.51	.928	1.182
14	3.53	14	.235	1.797	1.161	2.154	2.129	.31	.342	.690
		15	.142	2.412	1.753	3.915	2.278	.20	.281	1.522

Note. BILOG EX6 Data Set (n=1,000), CTT=classical test theory; Rasch= Rasch model; 1P= 1-parameter IRT model; 2P= 2-parameter IRT model; 3P= 3-parameter IRT model.
[a]The classical estimate is the number of correct answers.
[b]The classical estimate is the percentage of examinees correctly answering the item
[c]The classical estimate is the uncorrected item discrimination correlation coefficient.

Table 3
Comparability of Person Ability Statistics from the Two Measurement Frameworks: Correlations between CTT and Rasch Ability Statistics

	N	Ability
N	-	.989[a]
Person Ability		-

Note. Table represents estimates of individual abilities as reflected by the number of correct item responses.
[a]Correlation between the number of correct answers (N) and ability (θ)

Table 4
Comparability of Item Statistics from the Two Measurement Frameworks: Correlations between CTT-, Rasch-, and IRT-Based Item Difficulty indexes.

	CTT	Rasch	1P	2P	3P
CTT	-	.983[a]	.984	.939	.952
Rasch		-	.999	.966	.983
1P			-	.968	.983
2P				-	.978
3P					-

Note. CTT=classical test theory; Rasch= Rasch model; 1P= 1-parameter IRT model; 2P= 2-parameter IRT model; 3P= 3-parameter IRT model.
[a]Correlations between CTT item difficulty indexes with IRT item difficulty estimates derived from one- (Rasch also), two-, and three-parameter IRT models, respectively.

Table 5
Comparability of Item Statistics from the Two Measurement Frameworks: Correlations between CTT-, Rasch-, and IRT-Based Item Discrimination indexes.

	CTT	2P	3P
CTT	-	.841[a]	.510
2P		-	.584
3P			-

Note. CTT=classical test theory; 2P= 2-parameter IRT model; 3P= 3-parameter IRT model.
[a]Correlations between CTT item discrimination indexes with IRT item discrimination estimates derived from two- and three-parameter IRT models, respectively.

TM034290

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Classical test theory and item response theory: Analytical and empirical comparisons	
Author(s): Dae-Yeop Hwang	
Corporate Source: University of North Texas	Publication Date: February 14, 2002

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

Level 1

☒ Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2A

Level 2A

☐ Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only

The sample sticker shown below will be affixed to all Level 2B documents

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

2B

Level 2B

☐ Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, please

Signature: Dae Yeop Hwang	Printed Name/Position/Title: Dae-Yeop Hwang	
Organization/Address: University of North Texas	Telephone: (940) 565-4114	FAX:
	E-Mail Address: dyhwang@unt.edu	Date: 2-19-2002

(Over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
4483-A Forbes Boulevard
Lanham, Maryland 20706

Telephone: 301-552-4200
Toll Free: 800-799-3742
FAX: 301-552-4700
e-mail: ericfac@inet.ed.gov
WWW: http://ericfacility.org